DOCUMENT RESUME

ED 314 919 EC 222 079

AUTHOR Sutherland, Doris J.; And Others
TITLE An Examination of the Peer Review Process.
INSTITUTION Office of Special Education and Rehabilitative Services (ED), Washington, DC.
PUB DATE 89
NOTE 22p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Disabilities; Elementary Secondary Education; Federal Aid; Grants; *Improvement Programs; Outcomes of Education; *Peer Evaluation; *Professional Training; *Program Proposals; *Reliability
IDENTIFIERS Discretionary Programs

ABSTRACT

This study evaluated the effect of training on improving the reliability of the peer review process by determining whether or not training made a difference in the variability among reviewers' scores and documentation provided to support the scores. Different levels of training were provided for participants in the peer review process who were reviewing applications for discretionary grants in the Training Personnel for the Education of the Handicapped program. Reviewers' total scores, total word count for strengths and weaknesses, and approval/disapproval decisions were then examined. There were no significant differences in applications' average total scores across the three levels of training, although standard deviations were higher with more training. Also, there was no reduction in the dispersion of the scores with an increase in training. Reviewers who received additional training tended to write more words in both the strengths and weaknesses sections. Concerning approval or disapproval of the application, there were statistically significant differences across the training conditions, but they could not be determined to be training differences. Reviewers across all training conditions provided more documentation for those applications they disapproved than for those they approved. More experienced reviewers tended to give lower scores and more documentation than less experienced reviewers. (JDD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED314919

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

An Examination of the Peer Review Process

by

Doris J. Sutherland

M. Angele Thomas

Kathleen Hebbeler

Louis Danielson

Norman D. Howe

Office of Special Education and Rehabilitative Services
U.S. Department of Education
Washington, D.C.

1989

EC 222079

ABSTRACT

The purpose of this study was to identify ways to improve the quality of the Training Personnel for the Education of the Handicapped program (CFDA 84.029) review process by determining how much training would be beneficial to ensure that the reviewers have a similar orientation to the process. To do this, the study examined the degree of consistency among groups of reviewers who received various amounts of orientation training.

AN EXAMINATION OF THE PEER REVIEW PROCESS

A study of the effect of training to improve the reliability of the peer review process was conducted in 1988 using competitions of the Training Personnel for the Education of the Handicapped program, CFDA 84.029. This program is administered by the Division of Personnel Preparation (DPP) in the Office of Special Education and Rehabilitative Services (OSERS), United States Department of Education. The purpose of this discretionary grant program is to increase the quantity and improve the quality of personnel available to educate children and youth with disabilities. In fiscal year 1988, 230 reviewers were used to evaluate 839 new applications submitted to this program. Of the applications received, 244 were funded.

Although the size of the peer review panels and the approach to the process has varied since legislation was first enacted in 1958, the goal has remained constant: to select the best personnel training programs for Federal support based on written applications. The human and monetary cost of preparing an application and the number of applications competing for scarce Federal resources make it imperative that the process be continually scrutinized in order to determine possible improvements.

The study sought to determine whether or not training of reviewers made a difference in the variability between panelists' scores and the documentation provided to support the scores. Two groups of reviewers were given additional training. The result of training in those groups was compared with reviewers in a control group to determine what effect, if any, training had on increasing the number and quality of comments and improving the consistency between scores and comments.

Background

In 1982 the DPP staff developed a draft "Outline of a Plan to Improve the Quality of Personnel Preparation" (DPP, December 2, 1982) which sought to critique the current review process for possible improvements. Selected areas were prioritized for attention, i.e., providing a common frame of reference for applicants, reviewers and DPP staff by developing, with the field, indicators of quality for each selection criteria. Their work resulted in The Baseline Book (Smith-Davis, Morsink, & Wheatley, 1984) and the Quality Project Planning Document (DPP, 1983, 1984, & 1987).

DPP staff proceded with the next priority: the review process itself. Forms used by reviewers were redesigned to encourage more complete justification of scores given. A further streamlining step, taken in FY 1986, translated for reviewers how comments relate to scores - again providing a common frame of reference for all parties.

Work on these steps is reviewed annually, and as problems are identified, a process for addressing the issues is implemented. The new priority in 1988 was to examine procedures to improve the consistency of scores within panels and determine effective ways for encouraging reviewers to provide adequate documentation to justify their scores.

A review of the literature (1970 to 1988) revealed a limited number of studies concerning the peer review of discretionary grants. Several studies of peer review at the National Institute of Health and the National Science Foundation have been conducted that assessed various aspects of their approaches (Sanders, March 15, 1982), but no study was found that evaluated the consistency of reviewer feedback or how it might be increased.

In an attempt to improve quality, information was sought on effectiveness training frequently used by the business and industry. Training peer reviewers, if effective, would be in line with a primary concept promoted by philosophers and theorists of quality (Crosby, 1979, Deming, 1982, Juran, 1974). They advocate "prevention versus detection" [of an error] is a key strategy in improving quality which leads to productivity at cost gains. In addition, the prevention strategy builds credibility, trust, and ultimately, pride in a workforce dedicated to quality improvement (McCarthy, 1986). Juran (1986) contends that it is necessary to conduct training to assist personnel in carrying out appropriate preventive changes.

Translated into the peer review process, there is basis to believe that the training strategy would result in more consistent scores and thorough documentation for applicants, reviewers, and DPP staff. Clear and accurate peer reviewer feedback should increase the quality of submitted applications and have greater potential for impact on training programs. In addition, DPP would be able to redirect the time spent on responding to issues emanating from the peer review process to more substantive program work. Consequently, the Division of Personnel Preparation in fiscal year 1988 attempted to determine if specific training for the task of reviewing applications would result in more consistency among panel members and in recommendations being more adequately documented by reviewers for the applicant agency.

The technical review process of evaluating applications submitted for Federal funding under The Education of the Handicapped Act as amended by P.L. 99-457 (EHA) is generally followed across all programs within the Office of Special Education and Rehabilitative Services (OSERS).

Selection of Grants

Although extenuating circumstances may cause exceptions in either direction, typically the applications given higher scores have a greater the probability of being funded until all money set aside for that priority is awarded. In addition to agency-wide guidelines, DPP which administers Part D of EHA,

published Federal regulations in 1984 for the Training Personnel for the Education of the Handicapped program. The rules and regulations in existence at the time of the present study established criteria for selecting an application on the merits of its need, program content, plan of operation, evaluation plan, quality of key of personnel, extent of resources and budget. Thomas and Sutherland (1987) have described the various phases of how applications are selected.

Selection of Reviewers

The process of selecting reviewers in the discretionary grant programs is governed by the statute (Sec. 643), the Education Department General Administrative Regulations (EDGAR, Sec. 75.217), and a Department of Education Directive entitled "OSERS Enhanced III-2" (C-GPA:1-102). A total of 92 reviewers participated in the study. One-third of the reviewers were reading proposals for the first time, 20 percent were reading for the second time, and the remainder had paneled three or more times. Ninety-two percent held doctorates and eight percent had masters degrees.

Procedures for Reviewer Orientation

The procedures for orienting reviewers of personnel preparation applications from 1982-1987 consisted of showing a videotape which explained the OSERS guidelines followed by a discussion of the focus of the priority and the funding limitations. Subsequently, panelists checked their assigned

roster of applications for conflict of interest, and proceeded to read each application independently. After reading an application, each reviewer completed a scoring form giving the application a point value with respect to the evaluation criteria and documenting strengths and weaknesses. Two days later panels convened to discuss their individual deliberations and decide on a group recommendation for either approval or disapproval.

During fiscal year 1988 several modified types of orientation were employed. The purpose of this study was to identify for DPP and for applicants to the Training Personnel for the Education of the Handicapped program (CFDA 84.029) ways to improve the quality of the review process by determining how much training is necessary and beneficial to ensure that reviewers have a similar orientation to the process of evaluating applications. To do this, the study examined the degree of consistency among groups of reviewers who received various amounts of orientation. The findings of this study may also provide implications for other discretionary grants programs in the Office of Special Education Programs and in the Department of Education.

The questions to be investigated by this study were:

1. What degree of consistency exists among reviewers with the present level of training?
2. Would that degree of consistency be higher if more training were given?

3. How much more training would be beneficial to reviewers?

4. Do reviewers with more training provide more documentation to applicants?

Method

Three competitions which support masters and bachelors level training - the Preparation of Special Educators competition, the Preparation of Personnel to Work in Rural Areas competition, and the Preparation of Personnel for Minority Handicapped Children competition - were selected for this study because of the large numbers of similar applications that are submitted annually for consideration. Since the training programs supported by these competitions must be preservice, virtually all applicants are institutions of higher education. The applications in the three competitions describe training across all age groups, (0-21 years) and all content areas of special education. Three weeks of paneling applications from these competions resulted in a funding slate, which took into consideration the recommendations of reviewers, as well as DPP/OSERS recommendations to approve or disapprove applications.

The effect of minimal training on the consistency in scores among reviewers and the thoroughness of the documentation was determined by giving the traditional orientation to reviewers for the Rural and Minority competitions in May. Two other levels of training were used during the Special Educators competition in December. One single application was given to every reviewer who

participated in these competitions to determine the effect of training. Care was taken that reviewers were not made aware of the presence of a common application. The degree of similarity of scores among reviewers and the amount of documentation provided by the reviewers under each of the training conditions was examined. Although quality of documentation to support the scores given by panelists was an area of concern, no measure of quality could be used at this time. However, the number of words in the comments section were counted, as a proxy variable, to assess the quality of the review.

The three levels of training and the corresponding competitions were:

LEVEL I TRAINING: (RURAL AND MINORITY COMPETITIONS, May 9-12 and 16-19, 1988). The reviewers received the DPP Technical Application Review Document in the mail prior to coming to Washington. This was the current method of instructing reviewers. When they arrived, an orientation was held at which time reviewers were exposed to the information according to traditional practice. (Care was taken that the same information was given at each of the three sessions.)

LEVEL II TRAINING: (SPECIAL EDUCATORS COMPETITION, November 30-December 3, 1987). The reviewers were sent the same Technical Application Review Document in the mail. In addition, for training purposes an application from FY 1986 which is public information, was sent to them (with prior knowledge of the grantee). Each reviewer was requested to score it before the

orientation meeting. At that time, a discussion of the application took place. It was led by a demonstration panel. These experts were reviewers selected from among the panelists by the Division Director and the Branch Chiefs on the basis of their past performance as application evaluators. Scores were collected and analyzed from all reviewers upon their arrival so that they received feedback prior to beginning the actual review process.

LEVEL III TRAINING: (SPECIAL EDUCATORS COMPETITION, December 7-10, 1987). The third level training basically followed the same procedures as Level II. However, an additional level of training was implemented. A sampling of actual comments was distributed. The reviewers rated the comments from poor to excellent and determined characteristics of good comments.

In all cases, reviewers were instructed to not change their original scores, but were given the opportunity to adjust post-panel scores after the discussion. The intention was to gather information based on the data generated from the common application regarding how effective the written instructions were, how effective the demonstration panel discussion was, and how much change occurred as a result of the second and third levels of training.

Analysis

The impact of the training on the review process was

examined by looking at the reviewer's total score after paneling[1], the total word count for strengths and weaknesses, and the reviewer decision with regard to approval or disapproval. The common application was given an average score of 65. There were no significant differences with regard to the average total score across the three levels of training although the scores tended to get higher with more training. More importantly, there was no reduction in the dispersion of the scores with an increase in training; in fact, the standard deviations were slightly higher for those with more training. The data also show the equally wide spread of scores under each of the training conditions.

The variation in scores <u>within</u> panels was another way to examine the effect of training on scores. Analyzing panel means rather than individual scores reduces the dispersion somewhat, but there were no differences as a function of training.

With regard to word count and percent recommended for approval, reviewers wrote more words describing weaknesses than strengths. Reviewers who received additional training tended to write more words in both the strengths and weaknesses sections

---

[1] Reviewer's scores before and after the panel discussion were collected. Analyses of the two scores indicated that reviewers changed their scores very little (1.67 points on the average) as a result of the panel discussion. Because of the similarity in the two scores, the findings with regard to impact of training on total score are limited to the score after the panel discussion.

but these differences were not significant. Also, the standard data indicated that training did not reduce the variation among reviewers in the number of words written.

Overall, 23 percent of the reviewers recommended the application for approval. There were large, statistically significant differences across the training conditions. These differences, however, may reflect differences across competitions more than training differences. Examination of the reviewers' comments showed that the common application did not sufficiently address the focus of one of the competitions (which was not the competition to which the proposal was originally submitted). Far fewer reviewers in that competition (which was used for Level I) recommended the proposal for approval. However, Levels II and III were part of the same competition and still had very different approval rates. Discussion about some of the unintended differences across the three training sessions is provided in the "Implications" section.

One last way to look at the impact of training was to examine a plot of the data for the individual reviewers. It showed each reviewer's total score by the number of words written under the "Weakness" section. There were reviewers in all three training conditions who scored the proposal high, middle or low. Similarly, each condition had reviewers who wrote very little, an

average amount, or a great deal. Ideally, a "consistent reviewer" is one who, if he or she scores the application high,

has little to say about the proposal's weaknesses (although writing more never causes any problems). A similarly consistent reviewer is one who scores low but documents thoroughly the nature of the application's shortcomings. Both types of good reviewers were in all three training conditions. The "inconsistent reviewer" is the reviewer who scores the proposal low but writes little about what is wrong with it so that the unsuccessful applicant receives very little feedback and gains minimal knowledge from the review process. Again, this type of reviewer was found among all three training conditions.

Possibly, experienced reviewers are more consistent in their reviews than those who are new to this process. The review data were analyzed with respect to reviewer experience (first, second, or third time or more) These data show no consistent pattern in the score itself or the dispersion of scores as a function of reviewer status. Although there seemed to be a trend for the more experienced reviewers to write more, these differences were not statistically significant. (There were large standard deviations for words written among all levels of reviewer experience.)

To examine the issue of reviewer consistency, score and word count were examined with reference to approval rate. Reviewers recommending approval scored the application significantly higher

than those recommending disapproval. While there were no differences between the two groups in how much they wrote about

the application's strengths, those reviewers who recommended disapproval had significantly more to say about the application's weaknesses.

Several limitations need to be considered when reviewing the study. The complexion and configuration of panelists were not identical across the three levels with regard to experience in training personnel, expertise in content, dedication to task, attention to detail, and prior participation in the paneling process. The demonstration panels were not the same during Level II and III training which resulted in different discussions. There was a perception that one was noticeably better than the other. The panel environment, panel monitors and competition managers varied across training conditions which could have led to different results had these variables been controlled. The small number of reviewers in the study may prohibit drawing a definitive conclusion about the impact of training. Extraneous negative circumstances and events occurred during Level III training, e.g., outside interruptions during the training, and finally, all conclusions are based on the review of one application.

Implications

As previously stated, the primary goals of the training were to reduce variability in scores and to increase documentation. Since significant statistical differences were not found among the three training conditions, the experimental procedure may not have been a good test of the effect of training. The decreased

amount of time per application reviewed by panel members of Levels II and III may have interfered with any positive effect that resulted from the training since those reviewers were required to evaluate from two to three more applications than those in Level I. In addition, the time that the panelists in Level II and III had to actually complete their individual reviews was slightly lessened because of the training activities. However, that differential may have been offset by the additional training which was intended to better prepare reviewers for the task. The quality and length of the training exercise may have contributed to producing more discerning panelists in Level II. It is interesting that even with the additional workload and slightly reduced time, reviewers in Levels II and III did not significantly reduce documentation as would have been expected. This finding is enough to encourage the DPP staff to recommend assigning fewer applications per reviewer, offer additional training and then further evaluate the results.

Another interesting finding was that reviewers across all training conditions provided significantly more documentation for those applications they disapproved than for those they approved. It was also noted that more experienced reviewers tended to give lower scores than less experienced reviewers. Perhaps that is a result of new reviewers being less familiar with the quality of applications and the expectations of the panel groups.

There was little change in scores given by reviewers as a

result of discussing the application with other panel members. Thus the current process of convening panelists to discuss an application does not appear to reduce diversity in scores although feedback from the panel provides additional insight for applicant agencies. Consideration might be given to ways of changing dynamics to increase consistency, since divergent scores given to an application which has similar comments are very troublesome to applicants. Cole, Cole, and Reubin (Sanders, 1982) in their examination of the peer review process at the National Science Foundation, contend that the great bulk of reviewer disagreement"...is probably a result of real and legitimate differences of opinion among experts about what good science is or should be". If they are correct, this goal may be difficult to accomplish. Additional thought needs to be given to whether or not and, if so, how panelists should be encouraged to discuss applications and arrive at scores that are consistent among the panelists and with documentation that reflects the panel discussion.

Topics for Further Study

The examination of the effect of training on the peer review system has the potential to ultimately improve the process. Even though stated limitations of the study may have obscured the positive impact of training reviewer feedback was very supportive of the training. The following issues may be clarified through further studies:

1. Best methods of orienting reviewers to participate in the peer review process, e.g., time of orientation (before or during the orientation meeting, or both), content of the training (demonstration panel and/or quality of reviewers' comments);

2. Effective ways to encourage reviewers to document their scores adequately;

3. Efficient methods of evaluating the work of reviewers to ensure that applicants receive a high quality review; and

4. Other interventions that should be introduced to try to reduce variance in scores.

Conclusion

The Division of Personnel Preparation staff is concerned that the review process be as equitable and effective as possible and will continue to commit effort and energy toward its improvement. The Government must make funding decisions in a manner that is rational, intelligent and informed. This effort is facilitated by selecting reviewers who are highly qualified by virtue of their experience in the successful training of personnel and their outstanding program management. Over a hundred years ago, John Ruskin taught that quality is never an accident; it is always the result of intelligent effort.

The authors would like to express their appreciation to the Division of Personnel Preparation staff and especially to the three competition managers who assisted: Donald Blodgett, Victoria Ware, and Robert Gilmore. A special thanks to Lex Smith, a Doctoral Intern from University of Northern Colorado for his valuable contribution to this study.

REFERENCES

Crosby, P. B. (1979). Quality is free. New York: McGraw-Hill Book Company.

Deming, W. E. (1982). Quality, productivity, and competitive position. Cambridge, MA: Massachusetts Institute of Technology.

Division of Personnel Preparation, Office of Special Education Programs (1982). Outline of a plan to improve the quality of personnel preparation. Unpublished manuscript. Washington, DC: U.S. Department of Education.

Division of Personnel Preparation, Office of Special Education Programs (1983). Quality project planning document. Washington, DC: U.S. Department of Education.

Division of Personnel Preparation, Office of Special Education Programs (1984). Quality project planning document. Washington, DC: U.S. Department of Education.

Division of Personnel Preparation, Office of Special Education Programs (1987). Quality project planning document. Washington, DC: U.S. Department of Education.

Education Department General Administrative Regulations. 34 CFR Part 75 (1985).

Education of the Handicapped Act Amendments of 1986, §643, 20 U.S.C. §1443 (1987).

Juran, J.M. (Ed.) (1974). Quality control handbook (3rd ed.). New York: McGraw-Hill.

Juran, J.M. (1986). Quality progress, the quality trilogy. New York: McGraw-Hill.

McCarthy, M. (1986). Quality quake or trendy tremor. Performance management magazine. (9686-05) 1-22.

Office of Special Education and Rehabilitative Services (1987). Use of grant application reviewers. C:GPA: 1-102. Washington, DC: U.S. Department of Education.

Sanders, H.J. (1982). Peer review: How well is it working? Chemistry & Engineering, 3, 32-43.

Smith-Davis, J., Morsink, C., & Wheatley, W. (1984). The baseline book. Vienna, VA: Dissemin/Action.

Thomas, M.A. & Sutherland, D. (1987). The grant cycle: From program announcement to final decision. Vienna, VA: Dissemin/Action.

Training Personnel for the Education of the Handicapped: Final Regulations, 34 CFR Part 318, Federal Register, 49(314), 28371-28377, (1984).